# SDCI Net: Collaborative Research: An integrated study of datacenter networking and 100 GigE wide-area networking in support of distributed scientific computing

Zhengyang Liu

Oct 25, 2013

Supported by NSF Grant OCI-1127340

University of Virginia

# Outline

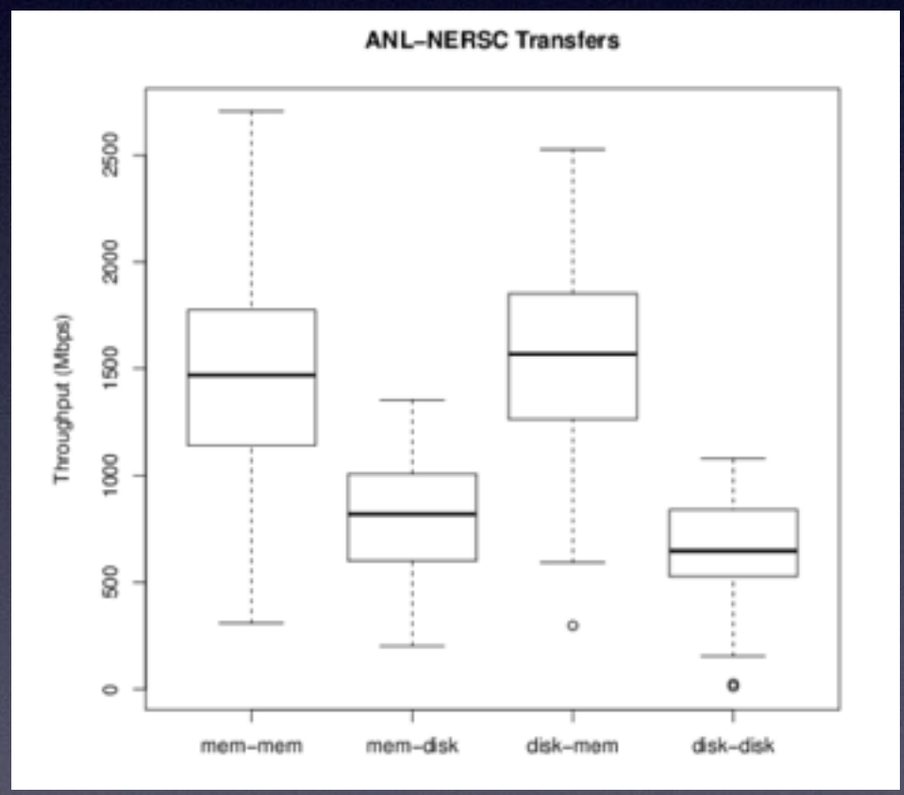
- Background
- Cause of Throughput Variance
- Experiment: Instrumented Transfers
- Regression Model
- Experiment: Disk I/O
- Future Work

# Throughput Variance

- 334 test transfers from ANL to NERSC
- 4 types: mem-mem (84), mem-disk (78), disk-mem(87), disk-disk(85)

TABLE VI: Throughput of ANL-NERSC transfers (Mbps)

| | mem-mem | mem-disk | disk-mem | disk-disk |
|---|---|---|---|---|
| Min | 308.9 | 202.4 | 297.4 | 10.85 |
| 1st Qu. | 1149 | 599.6 | 1265 | 527.3 |
| Median | 1472 | 819.0 | 1569 | 645.9 |
| Mean | 1463 | 789.6 | 1563 | 670 |
| 3rd Qu. | 1772 | 1007 | 1851 | 841.3 |
| Max | 2706 | 1354 | 2529 | 1079 |
| CV | 35.69% | 31.63% | 30.80% | 33.10% |




On using virtual circuits for GridFTP transfers, SC '12

# Problem Statement

1. Understand systematically the causes of variance and quantify the impact of each factor on throughput
2. Create validated models for transfer throughput as a function of these factors towards developing algorithms for determining amount of concomitant resource allocations

# Outline

- Background
- Cause of Throughput Variance
- Experiment: Instrumented Transfers
- Regression Model
- Experiment: Disk I/O
- Future Work

# Causes of Throughput Variance

- Intrinsic but less controllable factors
  - File size
  - RTT
  - Bottleneck link rate
- Intrinsic and controllable factors
  - Application and its parameters
    - e.g. “-p”, “-r” and “-fast” for GridFTP
  - Transport protocol and its parameters (e.g. TCP vs RDMA, TCP tuning)
- Extrinsic factors
  - Dynamic parameters of the end-to-end path (packet loss rate)
  - CPU/Memory
  - Disk

# Outline

- Background
- Cause of Throughput Variance
- Experiment: Instrumented Transfers
- Regression Model
- Experiment: Disk I/O
- Future Work

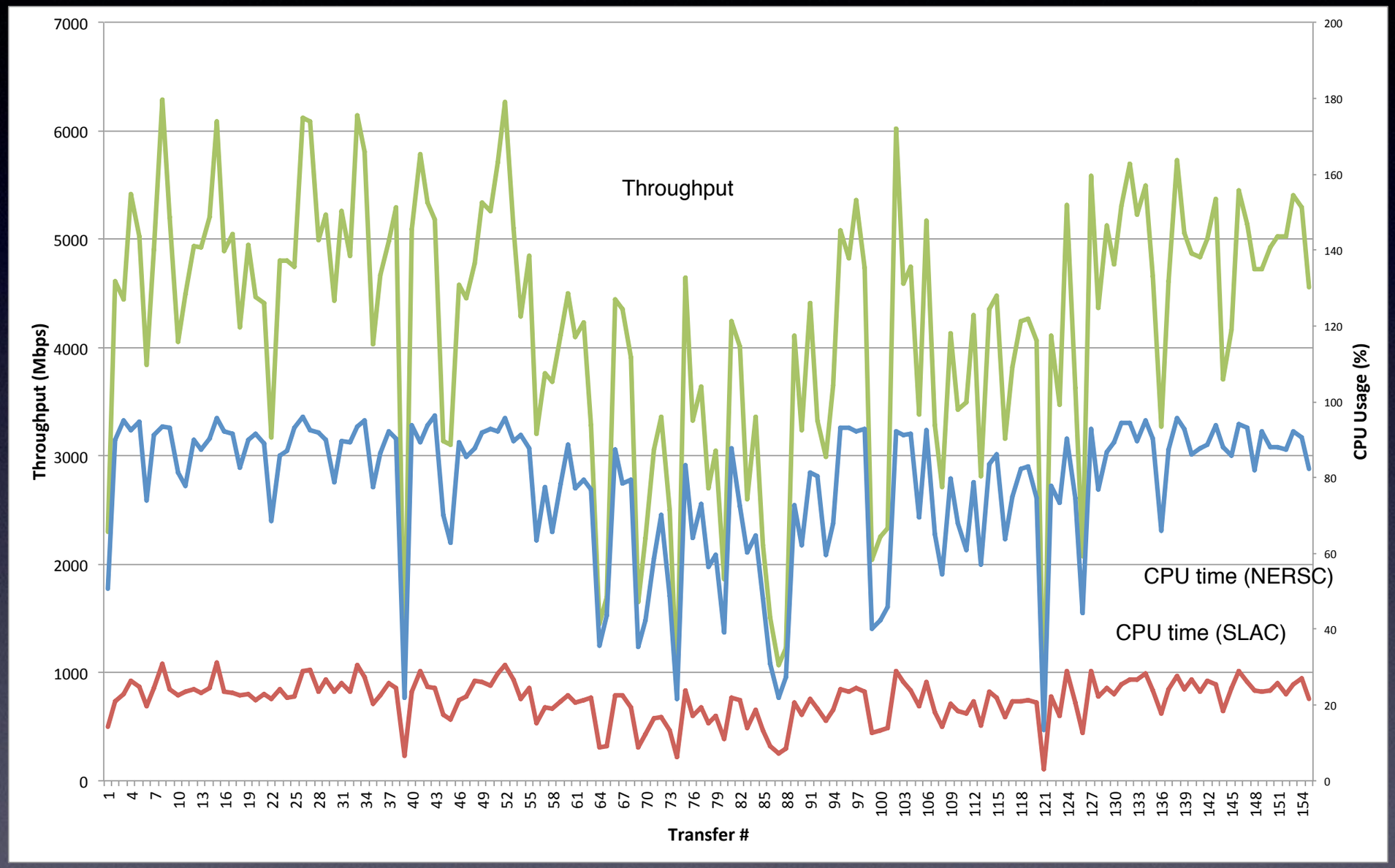
# Instrumented Transfers between production DTNs

- Monitoring scripts: initiates *top* (CPU usage) and *tcpdump* (packet loss) before each transfer
- Schedule hourly GridFTP transfers between NERSC and SLAC
- Analyze logs collected from GridFTP and monitoring tools

# Host Configurations

- NERSC DTN:
  - 2 x AMD Opteron (dual-core)
  - 8 GB
  - CentOS 5.8 (2.6.18 x86_64)
- SLAC DTN:
  - 2 x Xeon (quad-core with HT enabled)
  - 48 GB
  - RHEL 5.9 (2.6.18 x86_64)
- Bandwidth: 10 Gbps; RTT: 2.47ms (10 hops in between)

# Mem-mem Experiment




Transfer throughput vs CPU usage

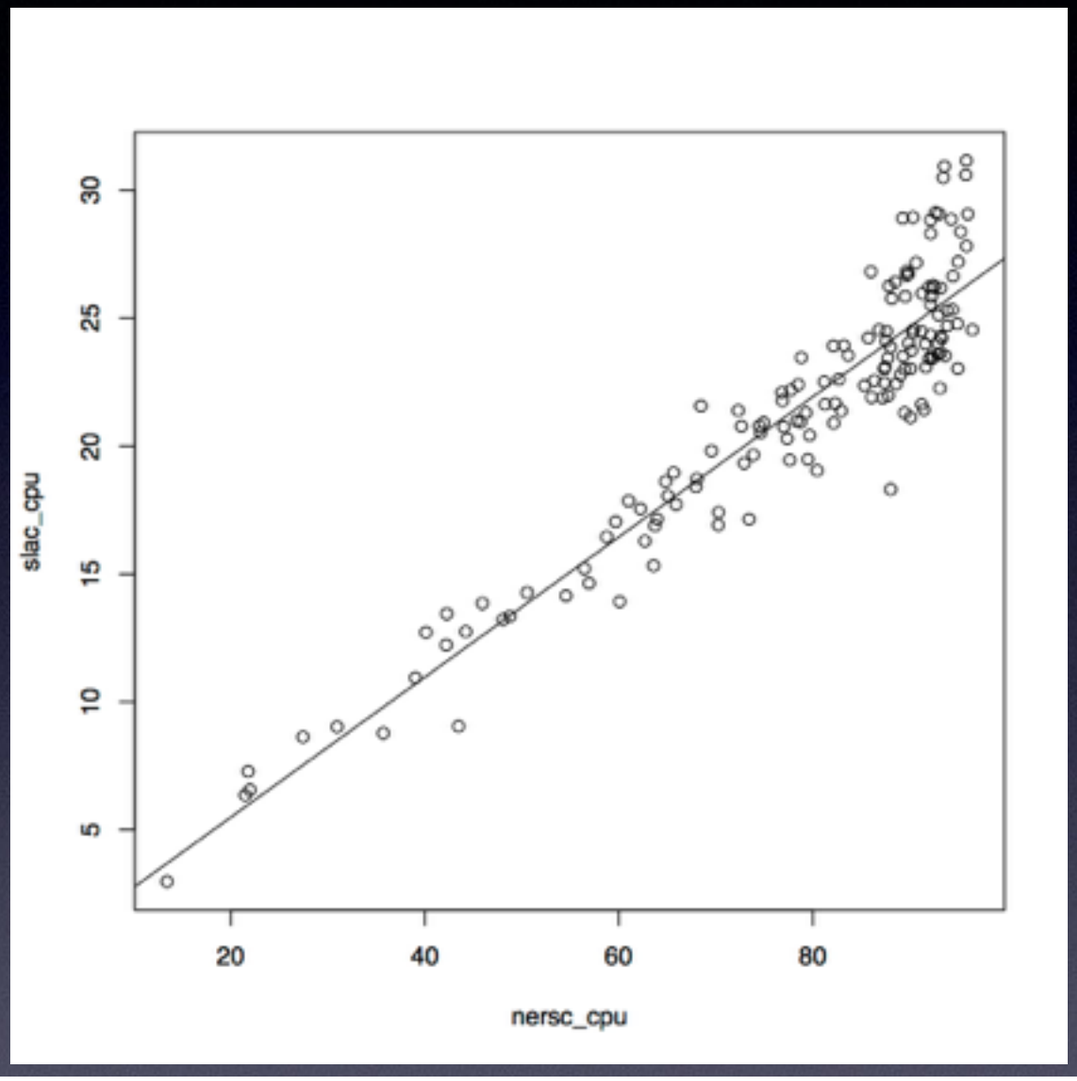
# Outline

- Background
- Cause of Throughput Variance
- Experiment: Instrumented Transfers
- Regression Model
- Experiment: Disk I/O
- Future Work

# Regression Model

- Dependent variable: throughput
- Independent variables:
  - NERSC CPU usage
  - packet loss rate
  - SLAC CPU usage?
    - SLAC CPU usage was highly correlated with NERSC CPU usage: linear model

# Regression Model




$$SLACcpu_i = \beta_0 + \beta_1 NERSCcpu_i + \epsilon_i \qquad (3)$$

# Regression Model

$$y_i = \beta_1' NERSCcpu_i + \beta_2' \epsilon_i + f(p_i) + e_i, \qquad (4)$$

Throughput

$p_i$: packet loss rate

recall Matthew Mathis equation:
$y_i \sim 1/\mathrm{sqrt}(p_i)$

*f* should be non-linear

use B-Spline

$$f(p_i) \approx \sum_{j=1}^{k+m} \alpha_j B_j(p_i) \qquad (5)$$

k = 2, m = 3 (chosen by LOOCV)

Acknowledgment: Prof. Jianhui Zhou from Department of Statistics taught us about B-spline and how to use it in R

# Regression Model

$\beta'_1 = 62.708, \beta'_2 = 153.194$

| Variable | Mean |
|---|---|
| Retransmission Rate ($p_i$) | 4.1E-05 |
| $f(p_i)$ in Mbps | -683.3 |
| Throughput (Gbps) | 4.226 |
| NERSC CPU usage (%) | 78.28 |
| SLAC CPU usage (%) | 21.45 |

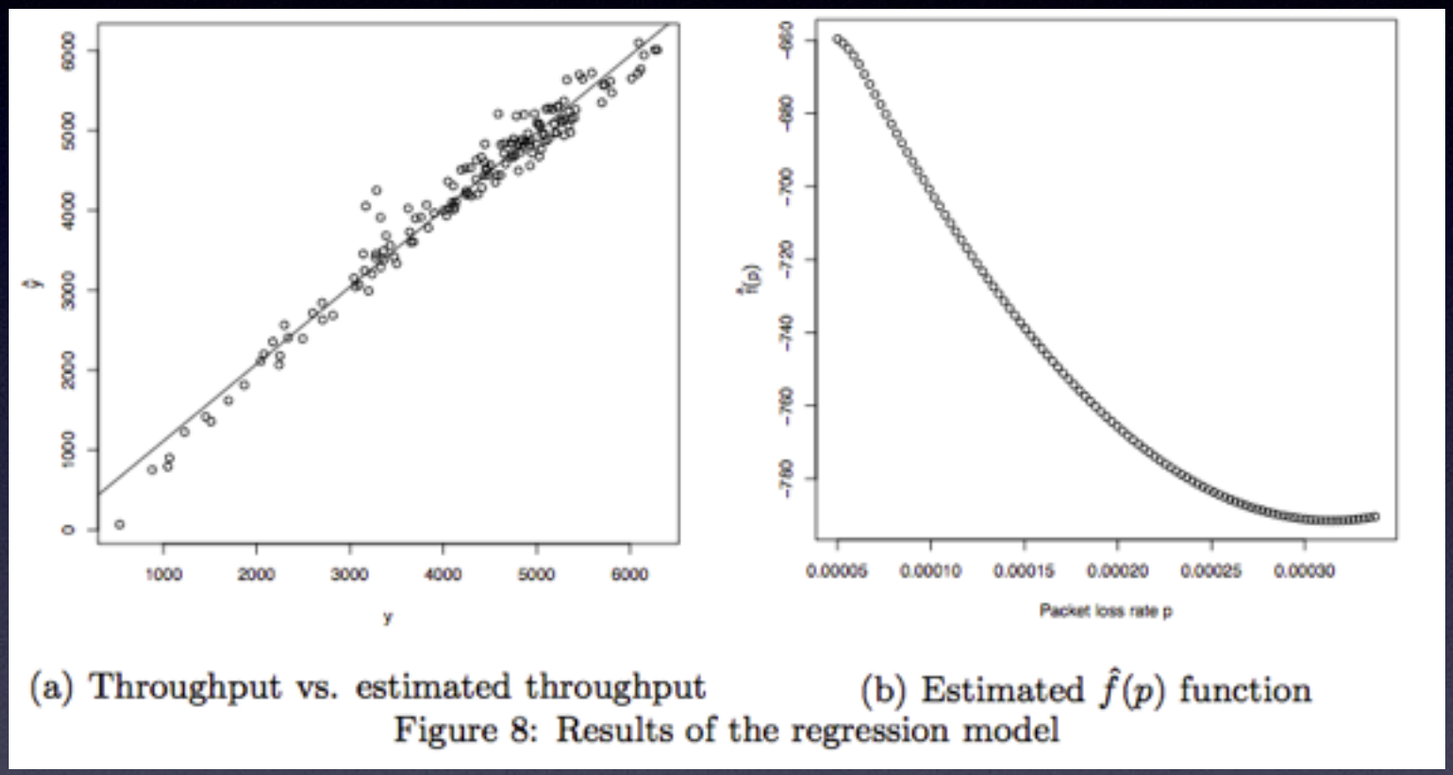
$$y_i = \beta'_1 NERSCcpu_i + \beta'_2 \epsilon_i + f(p_i) + e_i, \qquad (4)$$

Example: 62.708 * 78.28 + 153.194 * 0 - 683.3 = 4225.48 Mbps

## Observations

- CPU usage was the primary factor in determining throughput
- However packet loss rate, while small, also contributes to throughput (683.3/4225.48 = 16%)
- NERSC have slower CPUs than SLAC

# Regression Model




(a) Throughput vs. estimated throughput (b) Estimated $\hat{f}(p)$ function

Figure 8: Results of the regression model

Adjusted R-squared: 0.997

# Implications

- Computational nodes are shared in batch mode with a scheduler like PBS
- DTNs are used in interactive mode
- As users login to DTNs and initiate file transfer apps as needed, the amount of CPU and disk resources available to a particular transfer are not controlled
- To control variance, the number of concurrent process have to be controlled

# Outline

- Background
- Cause of Throughput Variance
- Experiment: Instrumented Transfers
- Regression Model
- Experiment: Disk I/O
- Future Work

# Disk-mem Experiment

1. Invoke dd to write an 8 GB file (first file) to the global scratch file system with the sync system call to force the completion of pending disk writes, and record the time taken for the write operation.
2. Invoke dd to write another 8 GB file (second file) to the global scratch file system to ensure that the first file is no longer in the filesystem cache, which is required for the next step.
3. Invoke dd to read back the first file (which is now on disk, not cache) and record the time taken for the read operation. (Conveniently second file will be forced out of the cache).
4. Using globus-url-copy, transfer the second file to /dev/null on the SLAC DTN, and use strace to record system calls for further analysis. The only disk I/O op- eration required is the reading of the second file, which we know is not in the cache.

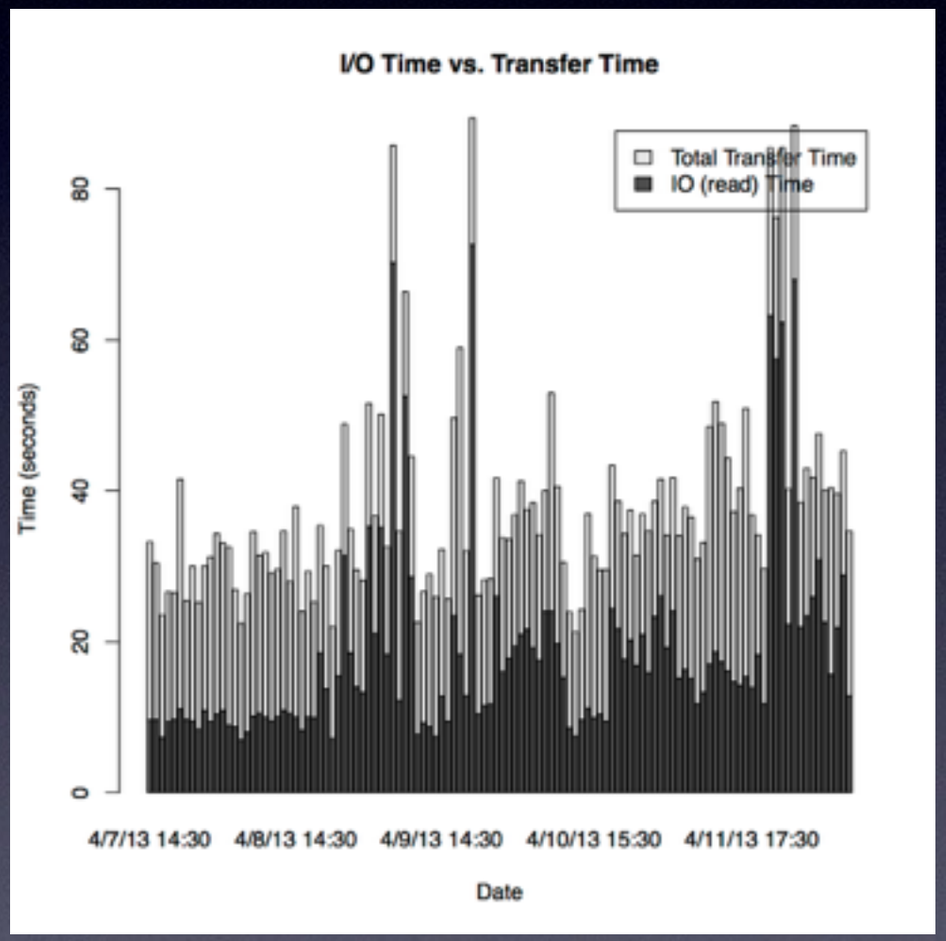
# Disk-mem Experiment

- "main loop" in globus-url-copy:
  - <starting up>
  - select
    - read
    - write
  - <clean up>
- total time spent in read measures disk I/O time

# Disk-mem Experiment (Preliminary)

Table 9: NERSC-to-SLAC disk-to-mem transfers

| | Ratio of Disk IO time to total transfer time | Throughput (Gb/s) |
|---|---|---|
| Min | 0.2647 | 0.768 |
| 1st Qu. | 0.3415 | 1.687 |
| Median | 0.4010 | 1.994 |
| Mean | 0.4524 | 1.989 |
| 3rd Qu. | 0.5523 | 2.32 |
| Max | 0.8191 | 3.228 |
| CV | 30.37% | 26.42% |



# Outline

- Background
- Cause of Throughput Variance
- Experiment: Instrumented Transfers
- Regression Model
- Experiment: Disk I/O
- Future Work

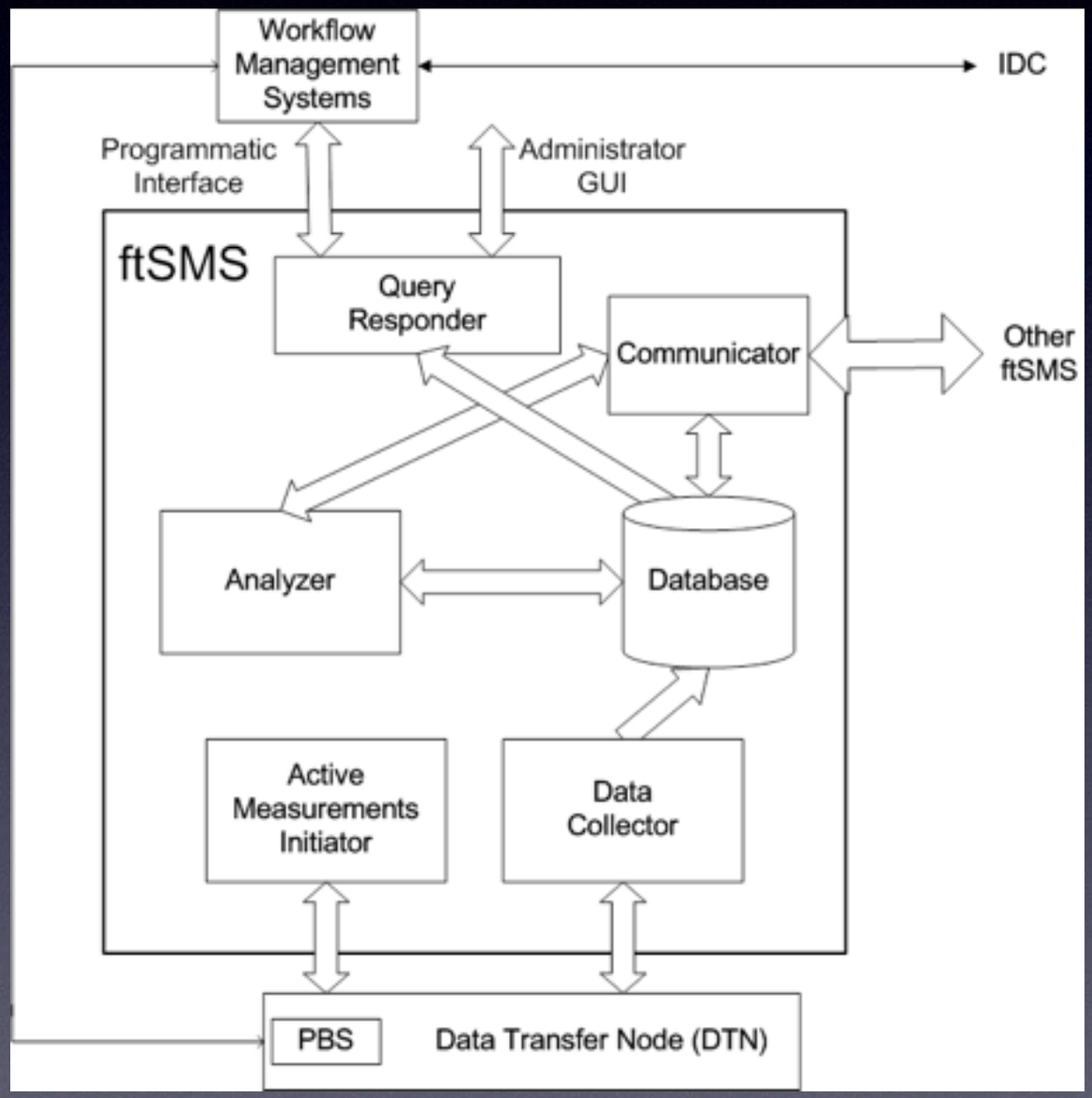
# Future Work

- develop statistical models including disk I/O and memory bandwidth
- prototype a file transfer service management system

# Thank you

UNIVERSITY of VIRGINIA

# File Transfer Service Management System



# B-Splines